RELATO DE
CAMPO · ARKHI

RELATO DE CAMPO · 01

O que a IA não conserta.

Relatos de campo como dicas de como adotar a IA no seu ambiente empresarial.

Por Marco Mendes

Arkhi — consultoria em estratégia e gestão da execução

arkhi.com.br

01 · ABERTURA

SINAL DE MERCADO

A McKinsey publicou em 2026 um dado que muda a conversa: empresas que reorganizaram processos em torno da IA já operam com 40% mais margem operacional que empresas que apenas a instalaram.

A pergunta deixou de ser "vamos adotar IA?". Passou a ser "estamos adotando a IA com abordagem de gestão de fluxo?".

Você está de carro novo mas o trânsito está engarrafado?

A IA dá poder ao seu time de desenvolvimento. É como se cada um deles estivesse de carro novo, muito mais possante. Mas mesmo assim isso não será útil para a sua eficiência operacional se os gargalos e restrições entre áreas e departamentos não for também melhorado, seja por processos e pela própria IA de alta maturidade.

A adoção de IA está acontecendo em ritmo acelerado. As ferramentas estão melhores. Os times estão mais letrados. Mas, quando olhamos para os indicadores que importam — tempo de entrega, custo do atraso, satisfação do cliente, vazão de ponta a ponta — o ponteiro praticamente não sem moveu.

Este material é um relato do que estamos observando. São padrões recorrentes que aparecem em empresas diferentes, de portes diferentes, em setores diferentes — e todos compartilham uma mesma raiz: a IA foi aplicada sobre processos que ninguém mediu antes.

A tese central deste documento é simples.

A IA **amplifica** o que já existe na sua operação. Em processos bons, ela acelera valor. Em processos ruins, ela acelera **desperdício**.

02 · O RETRATO DO MERCADO

A maioria está acelerando o indivíduo. Quase ninguém está acelerando a organização.

Existe uma forma de classificar a maturidade de adoção de IA que ajuda a explicar o que estamos vendo. Usamos as siglas ML (Maturity Level), para designar o escopo de atuação da sua IA.

ML0 · INDIVÍDUO

IA no nível do colaborador

Indivíduo: IA acelera o trabalho do colaborador em suas tarefas diárias.

ML1 · EQUIPE

IA no nível da equipe

IA acelera o enlace entre pessoas do mesmo time, reduz retrabalhos nos fluxos de entrega e começar a acelerar a geração de entregáveis.

META

ML2/ML3 · FLUXOS / SERVIÇOS

IA em fluxos e sistema de trabalho

A IA identifica e reduz gargalos, filas e esperas entre equipes e departamentos. Ela contribui de forma efetiva para reduzir o tempo de entrega, aumentar a produtividade , acelerar e criar previsibilidade estatística de entrega.

MATURIDADE DE ADOÇÃO →

A distância entre uma camada é de gestão. A empresa que está no ML0 não chega aos níveis de Fluxo e Serviços apenas adotando as últimas versões de LLM da Anthropic, OpenAI ou Google. Ela chega entendendo, de forma quantitativa, como o trabalho flui — e onde, exatamente, ele para de fluir. E aplica a IA para liberar essas restrições.

— 03 · A MATEMÁTICA QUE NINGUÉM FAZ

80% do trabalho está parado. Dobrar a velocidade dos 20% restantes muda muito pouco.

EFICIÊNCIA DE FLUXO TÍPICA EM TRABALHO DO CONHECIMENTO

TEMPO DE ENTREGA TOTAL

<20% EFETIVOS

Em ambientes de trabalho do conhecimento, a maior parte do tempo de uma demanda não é tempo de execução. É tempo de espera.

Espera por revisão. Espera por aprovação. Espera por dependência de outra área. Espera por homologação. Espera por priorização. Espera por contexto. Espera porque alguém está em férias. Espera porque alguém está fazendo outra coisa.

Quando medimos isso em sistemas reais, o padrão se repete: o trabalho passa em média **80% do tempo em filas** e apenas **20% em execução efetiva**.

Agora, a matemática.

Se o trabalho está 80% parado e 20% em execução, tornar a parte de execução duas vezes mais rápida com IA melhora o tempo total de entrega em apenas 10%. Não importa quão bom é o modelo. Não importa quantas licenças foram compradas. Se a empresa não tocou nos 80% em fila, o ponteiro não se move.

E há um agravante. Quando a IA acelera a execução sem que ninguém olhe para a fila, a fila **cresce**. Times geram mais documentos ou códigos — e empurram para a próxima etapa um volume maior do que ela consegue absorver. Fila maior do que antes, com a sensação enganosa de que o trabalho está fluindo.

UM EXEMPLO — COMO VOCÊ MEDE O CUSTO DA SUA IA?

O custo por token é um exemplo ML0. Mas é melhor medir o custo por tarefa concluída — métrica que conecta IA diretamente ao resultado de negócio.

O custo do talento é alto. O custo do atraso é maior.

04 · RELATO 1

01 A POC de fim de semana.

Um caso real.

O QUE ACONTECEU

Um gestor sênior assinou, por conta própria, a licença de uma ferramenta de codificação assistida por IA. Passou o fim de semana inteiro testando, animado com o que conseguiu produzir sozinho. Na segunda-feira, levou uma POC bem feita para a reunião de liderança, conectando o experimento a um caso real da empresa. A decisão foi rápida: aprovar o piloto, alocar um time dedicado, expandir o uso para o desenvolvimento de produto. A expectativa é que em 1 semana a equipe entregaria um novo módulo do produto.

Dois meses depois, o time alocado ainda não tinha entregado resultados reais de negócio. A POC do fim de semana não escalou.

POR QUE ACONTECEU

A POC funcionou no fim de semana porque era um problema isolado, em um ambiente controlado, com uma única pessoa decidindo o que era bom o suficiente. Quando entrou no time, encontrou o que a POC não tinha: dependências de aprovações, dependência entre áreas, fila de homologação, regras de qualidade, contexto de produto, coordenação entre pessoas, reuniões para qualificar requisitos, testes. Não havia processo estruturado para integrar IA ao fluxo. Não havia linha de base de produtividade medida antes do experimento. Não havia critério explícito do que seria considerado sucesso.

APRENDIZADOS

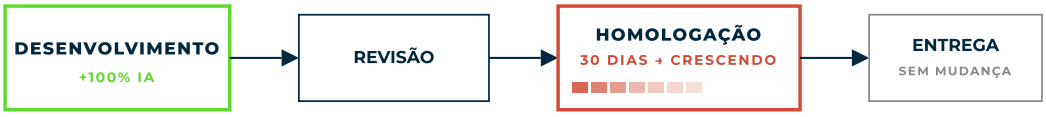
POCs individuais são positivos. Geram aprendizados e sempre úteis. O cuidado é projetar o que o indivíduo consegue fazer sozinho com alta eficiência de fluxo para um ambiente engarrafado com baixa eficiência de fluxo. IA ML0 não escala para ML3.

— 05 · RELATO 2

02 A métrica que mente.

Um caso real.

OTIMIZAÇÃO LOCAL × FLUXO GLOBAL



LEAD TIME TOTAL PARA O CLIENTE FINAL: INALTERADO

O QUE ACONTECEU

A empresa estabeleceu, como meta corporativa de adoção de IA, uma redução de custos baseada em **quantidade de linhas de código geradas**. A meta foi cumprida. A IA gerou volume. A liderança apresentou o número de horas economizadas. O caso virou referência interna. E o tempo de entrega para o cliente final continuou exatamente o mesmo.

POR QUE ACONTECEU

Quando a IA dobrou a produtividade na fase de desenvolvimento, a fila na homologação cresceu ainda mais. Mais código pronto, mais código esperando. E para piorar, o tempo de ciclo da etapa de homologação é de quase 30 dias. O que houve? O custo total do atraso subiu.

APRENDIZADOS

A métrica de linhas de código é uma **métrica local** — desconectada do resultado finalístico do negócio. Toda iniciativa de IA precisa estar conectada a métricas que importam para o cliente final. Quando a meta é eficiência local, o sistema vai entregar eficiência local — e ineficiência global. Áreas que podem brilhar individualmente, mas com fluxo travado coletivamente.

O problema não está em quem toca, mas sim em ninguém ouvir o todo.

06 · RELATO 3

03 A delegação da compreensão.

Um caso real.

O QUE ACONTECEU

Um projeto de modernização adotou IA para acelerar a geração de código desde o início. A IA gerou velocidade na superfície. Mas as equipes não estavam preparadas para usar a IA em escala. Não havia processos, governança e mecanismos de validação do código produzido. Com o crescimento do produto, a IA não conseguiu mais operar e começa a "alucinar" com mais frequência. A taxa de defeitos subiu e os POs foram muito críticos com a baixa qualidade do produto.

POR QUE ACONTECEU

A IA foi aplicada como substituta do pensamento, não como amplificadora dele. O código foi gerado, integrado, entregue — e nunca verdadeiramente entendido pelo time que precisaria mantê-lo. A crença ingênua é que a IA iria fazer código de ponta a ponta, sem intervenção humana mínima.

APRENDIZADO

Você pode delegar parte do aprendizado para a IA. Mas não pode delegar a compreensão do que é produzido.

07 · O QUE O MERCADO JÁ RECONHECE

Não é só a Arkhi dizendo isso.

No AI Festival 2026, líderes de Anthropic, McKinsey, Amazon e StartSe convergiram em um ponto: a IA não falha por inteligência. Falha por sistema.

Os três relatos das páginas anteriores poderiam ser lidos como casos isolados. Não são. Em maio de 2026, o maior encontro de IA do mercado brasileiro consolidou o mesmo diagnóstico que estamos descrevendo aqui, com palestrantes que operam em escalas que poucas empresas no mundo conseguem.

RAFAEL SIQUEIRA · MCKINSEY

GAP DE PRODUTIVIDADE

"A diferença entre empresas que apenas 'usam' IA e as que 'reorganizaram processos' em torno da IA já chega a 40% em margem operacional."

Leitura: É o argumento mais direto para justificar investimento e reorganização interna. O gap não é projeção — é dado de mercado.

HENRIQUE SAVELLI · ANTHROPIC

CUSTO POR TAREFA CONCLUÍDA

"O custo por token virou irrelevante. O novo padrão de avaliação empresarial deve ser o custo por tarefa concluída — métrica que conecta IA diretamente ao resultado de negócio."

Leitura: A linguagem do board mudou. Medições de produtividade por volume de output viraram obsoletas.

FELIPE BLANES · AMAZON AGI LABS

CONFIABILIDADE > INTELIGÊNCIA

"O que mudou nos últimos 12 meses não foi a inteligência dos modelos. Foi a confiabilidade. Modelos consistentes agora permitem colocar agentes em fluxos críticos sem medo de alucinações."

Leitura: O salto de 2026 não é técnico. É de previsibilidade. E previsibilidade só existe sobre sistemas que conhecemos.

CRISTIANO KRUEL · STARTSE

A BARREIRA PSICOLÓGICA

"A maior barreira para a IA em 2026 não é tecnológica. É psicológica. A liderança que insiste no comando manual em um mundo de autonomia agêntica tem o mesmo destino das empresas que ignoraram a internet em 1995."

Leitura: A obsolescência não vem por falta de ferramenta. Vem por falta de capacidade de orquestrar.

Esses quatro recortes não são sobre tecnologia. São sobre **gestão**. E é exatamente onde a Arkhi atua há mais de 16 anos: tornando o fluxo visível, medindo antes de intervir, e aplicando ferramentas — agora também a IA — nos pontos de restrição reais do sistema.

— 08 · O CONTRAPONTO

Quando funciona, funciona por motivo concreto.

Exemplos públicos de aplicação bem sucedida em IA.

NETFLIX

Algoritmo de recomendação como motor de negócio

EXEMPLO DE MÉTRICA-ALVO

Tempo de engajamento · Retenção · Redução de churn

ONDE A IA OPERA

Personalização profunda conectada a objetivos finalísticos do negócio

A IA não foi adotada porque era moda. Foi adotada para resolver um problema que a Netflix sabia descrever em termos quantitativos.

IFOOD

IA aplicada à restrição real da operação

EXEMPLO DE MÉTRICA-ALVO

Tempo de entrega de pedidos

ESCALA ATUAL

Mais de 10 mil agentes ativos em operação

— Isabella Piratininga, AI Festival 2026

ONDE A IA OPERA

Restrições estruturais que definem a experiência do cliente final

Não é IA aplicada à produtividade de quem programa o app. É IA aplicada às restrições reais da operação.

O que os dois casos têm em comum.

01

Entendimento prévio do fluxo de valor

Sabe os pontos de restrição. Mediu a linha de base antes da intervenção.

02

IA aplicada na restrição certa

Não no lugar mais fácil ou mais visível. No lugar que importa para o cliente.

03

Métrica finalística

O cliente final percebe. O negócio sustenta. Não é métrica de etapa.

09 · A TESE

A IA precisa de um sistema de trabalho para entregar resultado.

Os relatos das páginas anteriores não são sobre IA. São sobre o que acontece quando a IA é introduzida em sistemas de trabalho que ninguém mediu antes. A ferramenta é nova, mas o problema é antigo.

Há mais de duas décadas, a teoria das restrições, o pensamento de fluxo e o método KMM descrevem o mesmo padrão: organizações de trabalho do conhecimento não falham por falta de talento, mas por desconhecimento do próprio fluxo. Sem essa consciência, qualquer ferramenta nova — IA, automação, metodologia, plataforma — se transforma em mais uma camada sobre o caos invisível.

Drew Boyd e Jacob Goldenberg contrariam o conselho mais repetido do mundo dos negócios. Não se trata de pensar fora da caixa. Trata-se de pensar dentro da caixa — com as restrições reais do contexto, dos recursos, do mercado, da operação. É dentro dessas restrições que a inovação útil acontece.

Para a Arkhi, a leitura é direta. Três movimentos:

01

Tornar o fluxo visível.

Antes de qualquer ferramenta. Antes de qualquer modelo. Mapear como o trabalho realmente acontece — não como o organograma promete. Plataformas como o BusinessMap fazem parte do nosso ferramental para essa visualização.

↓

02

Medir antes de intervir.

Eficiência de fluxo, lead time, custo do atraso, throughput. Sem linha de base, não há como avaliar se a IA entregou o que prometeu — e qualquer narrativa de sucesso fica refém do entusiasmo da liderança.

↓

03

Aplicar IA nas restrições, não nas facilidades.

O lugar onde a IA mais entusiasma é raramente o lugar onde ela mais resolve. A pergunta certa não é *onde posso usar IA?*. É *qual restrição do meu sistema de trabalho a IA é capaz de aliviar?*

É nessa sequência que a IA deixa de ser produtividade individual e passa a ser resultado de negócio. É nesse trabalho que a Arkhi atua.

10 · GLOSSÁRIO + FECHAMENTO

Conceitos-chave, uma última pergunta e onde nos encontrar.

GLOSSÁRIO RÁPIDO

Eficiência de fluxo

Percentual do tempo total de uma demanda em que ela está sendo efetivamente trabalhada, em relação ao tempo total em que permanece no sistema. Em trabalho do conhecimento, raramente passa de 15%.

Custo do atraso

O custo financeiro, competitivo ou estratégico de uma demanda não ser entregue no momento em que entrega valor máximo. Em serviços, normalmente maior que o custo do talento — e raramente medido.

Custo por tarefa concluída

Métrica de avaliação de IA empresarial que substitui o "custo por token". Conecta IA diretamente ao resultado de negócio: o que importa não é quanto a IA gerou, mas quanto trabalho de fato foi concluído com o auxílio dela. Conceito sintetizado por Henrique Savelli (Anthropic) no AI Festival 2026.

KPI Fit For Purpose (F4P)

Indicador conectado à percepção real de valor do cliente final, não a métricas internas de etapa. Distingue medições que importam das que apenas alimentam relatórios.

Human-in-the-loop

Modelo de adoção de IA em que o humano permanece responsável pela validação e direcionamento. A pergunta crítica não é se o humano está no loop — é onde no loop ele precisa estar.

ML0 / ML1 / ML2 / ML3

Camadas de maturidade de aplicação de IA. ML0 é o uso individual. ML1 introduz a IA na coordenação entre pessoas do mesmo time, reduzindo retrabalho. ML2/ML3 são aplicações no nível de fluxos e serviços — onde a IA identifica e reduz gargalos, filas e esperas entre equipes, conectando-se ao fluxo de valor e aos KPIs finalísticos.

A PERGUNTA QUE FECHA

Minha IA está melhorando minha eficiência de fluxo? E meus resultados de negócio?

Se a resposta for *não sei*, a IA ainda não está entregando o que promete. E o problema não está na ferramenta.

A Arkhi é uma consultoria de aumento de eficiência de serviços. Trabalhamos com lideranças que precisam transformar direcionamento estratégico em resultado mensurável — e que reconhecem, ao olhar para o próprio sistema de trabalho, que existe valor represado esperando para fluir. Se este relato fez sentido para o seu contexto, vamos conversar.

arkhi.com.br • **wa.me/5561991443379**